11.Number of kids prediction - Preprocessing Data

- **This is the collected data**

Number of Kids	**Working Expereince (years)**	**Age**	**Salary**	**Blood Types**
3	15	45	$250,000	A
1	5	30	$200,000	B
2	10	38	$150,000	AB
1	<missing>	36	$180,000	O

1.Please clean the missing data using median approach.

- Pandas' Approach to replace the missing values of only one attribute: DataFrame’s dropna(), drop(), and fillna() methods.
- Noticed that the Working Expereince attribute has some missing value, you can fix the issue by using DataFrame’s dropna(), drop(), and fillna() methods:

```
# Option 1: Get rid of the corresponding districts.
data.dropna(subset=["Working_Expereince"])
# Option 2: Get rid of the whole attribute.
data.drop("Working_Expereince", axis=1)
# Option 3: Set the missing values to some value (zero, #  the mean, the median, etc.) on the training  set
median = data["Working_Expereince"].median()
data["Working_Expereince"].fillna(median, inplace=True)
```

- The **mean** is the sum of all the numbers in the set (30) divided by the amount of numbers in the set (3) ==> **30/3 = 10.**
- The **median** is the middle point of a number set, in which half the numbers are above the median and half are below. In our Example the median = **10.**
- In the **case** of a high number of outliers in your dataset, it is recommended to use the **median instead** of the **mean**. Another common method that works for both numerical and nominal features use the most frequent **value** in the column to **replace** the **missing values**.

Number of Kids	**Working Expereince (years)**	**Age**	**Salary**	**Blood Types**
3	15	45	$250,000	A
1	5	30	$200,000	B
2	10	38	$150,000	AB
1	10	36	$180,000	O

- **Sciki-Learn's Approach to replace missing values of all attributes: Imputer.**

Step 1: Create an Imputer instance, specifying that you want to replace each attribute's missing values with the median of that attribute:

Step 2: Since the median can only be computed on numerical attributes, we need to create a copy of the data .

Step 3: Fit the imputer instance to the training data using the fit() method.

Step 4: The imputer has simply computed the median of each attribute and stored the result in its statistics_ instance variable.

- **Only the Working Experience attribute had missing values, but we cannot be sure that there won't be any missing values in new data after the system goes live, so it is safer to apply the imputer to all the numerical attributes:**

```
imputer.statistics_
```

Step 5: Use this "trained" imputer to transform the training set by replacing missing values by the learned medians:

Program -

```
import numpy as np
from sklearn.impute import SimpleImputer
imp= SimpleImputer(missing_values=np.nan, strategy='median')
X = [[15], [5],[10],[np.nan]]
imp.fit(X)
```

print("statistics_ = ", imp.statistics_)

print("median = \n",imp.transform(X)) #median

Output -

statistics_ = [10.]

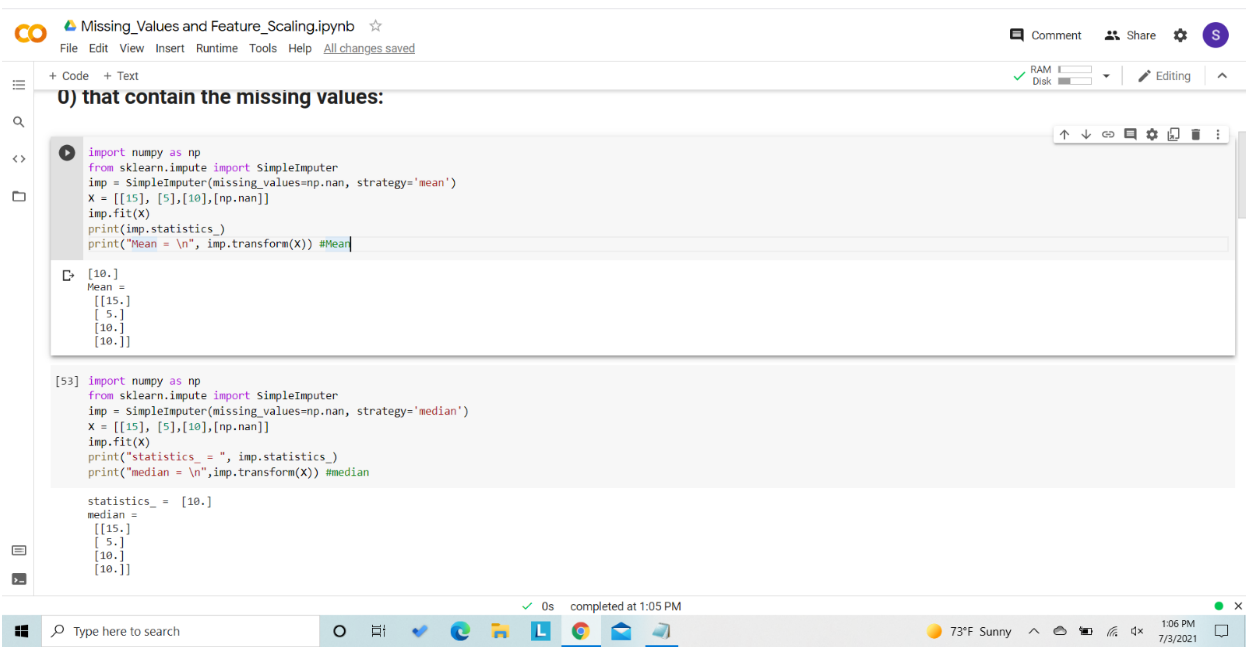
median =

[[15.]

[5.]

[10.]

[10.]] //Missing Value – Median



2.Please use Correlation to determine which of the following attributes is more related to "Number of Kids"?

- **Working Experience**
- **Age**

The correlation coefficient that indicates the strength of the relationship between two variables can be found using the following formula:

Definition

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Where:

- r_{xy} – the correlation coefficient of the linear relationship between the variables x and y.
- xi – the values of the x-variable in a sample.
- $\bar{x}$– the mean of the values of the x-variable.
- yi – the values of the y-variable in a sample
- $\bar{y}$ – the mean of the values of the y-variable

In order to calculate the correlation coefficient using the formula above, you must undertake the following steps:

- Obtain a data sample with the values of x-variable and y-variable.
- Calculate the means (averages) $\bar{x}$ for the x-variable and $\bar{y}$ for the y-variable.
- For the x-variable, subtract the mean from each value of the x-variable. Do the same for the y-variable.
- Multiply each (xi **-** $\bar{x}$**)** value by the corresponding b-value and find the sum of these multiplications (the final value is the numerator in the formula).
- Square each (xi - $\bar{x}$) value and calculate the sum of the result
- Find the square root of the value obtained in the previous step (this is the denominator in the formula).
- Divide the value obtained in step 4 by the value obtained in step 7.

- **Find the correlation coefficient of the linear relationship between the variables x (Number of Kids) and y (Working Experience).**

Number of Kids -> xi	**Working Experience (years) -> yi**	**xi - $\bar{x}$**	**yi - $\bar{y}$**	**(xi - $\bar{x}$) * (yi - $\bar{y}$)**	**(xi - $\bar{x}$)^2**	**(yi - $\bar{y}$) ^2**
3	15	1.25	5	6.25	1.5625	25
1	5	-0.75	-5	3.75	0.5625	25
2	10	0.25	0	0	0.0625	0
1	10	-0.75	0	0	0.5625	0

Mean($\bar{x}$) 1.75	**Mean($\bar{y}$) = 10**	**Sum**	**10**	**2.75**	**50**

Using the obtained numbers, can calculate the coefficient between the variables x (**Number of Kids**) and y **(Working Experience)** is :

r_{xy} = 10 / sqrt(2.75*50) = 0.85280286542

- **Find the correlation coefficient of the linear relationship between the variables x(Number of Kids) and y (Age).**

Number of Kids -> xi	**Age -> yi**	**xi - $\bar{x}$**	**yi - $\bar{y}$**	**(xi - $\bar{x}$) * (yi - $\bar{y}$)**	**(xi - $\bar{x}$)^2**	**(yi - $\bar{y}$) ^2**
3	45	1.25	7.75	9.6875	1.5625	60.06
1	30	-0.75	-7.25	5.4375	0.5625	52.5625
2	38	0.25	0.75	0.1875	0.0625	0.5625
1	36	-0.75	-1.25	0.9375	0.5625	1.5625
Mean($\bar{x}$) = 1.75	**Mean($\bar{y}$) = 37.25**	**Sum**		**16.25**	**2.75**	**114.75**

Using the obtained numbers, can calculate the coefficient between the variables x(**Number of Kids**) and y (**Age**) :

r_{xy} = 16.25 / sqrt(2.75*114.75) = 0.91476738366

The coefficient indicates that the following attributes “Working Experience” and “Age” have a high positive correlation. This means that their respective chances that the both attributes belong to the Number of Kids. Therefore, compare to both attributes, in fact, **Age attribute is more related to "Number of Kids".**

3. Please use One-Hot Vectors approach to convert the Blood Types.

Convert text categorical attribute “Blood Types” to be able to compute its median.

Step 1: Convert from text categories to integer categories

Note: One issue with Categorical Value representation is that ML algorithms will assume that two nearby values are more similar than two distant values.

Step 2: Convert from integer categories to One-Hot Vectors.

For example -

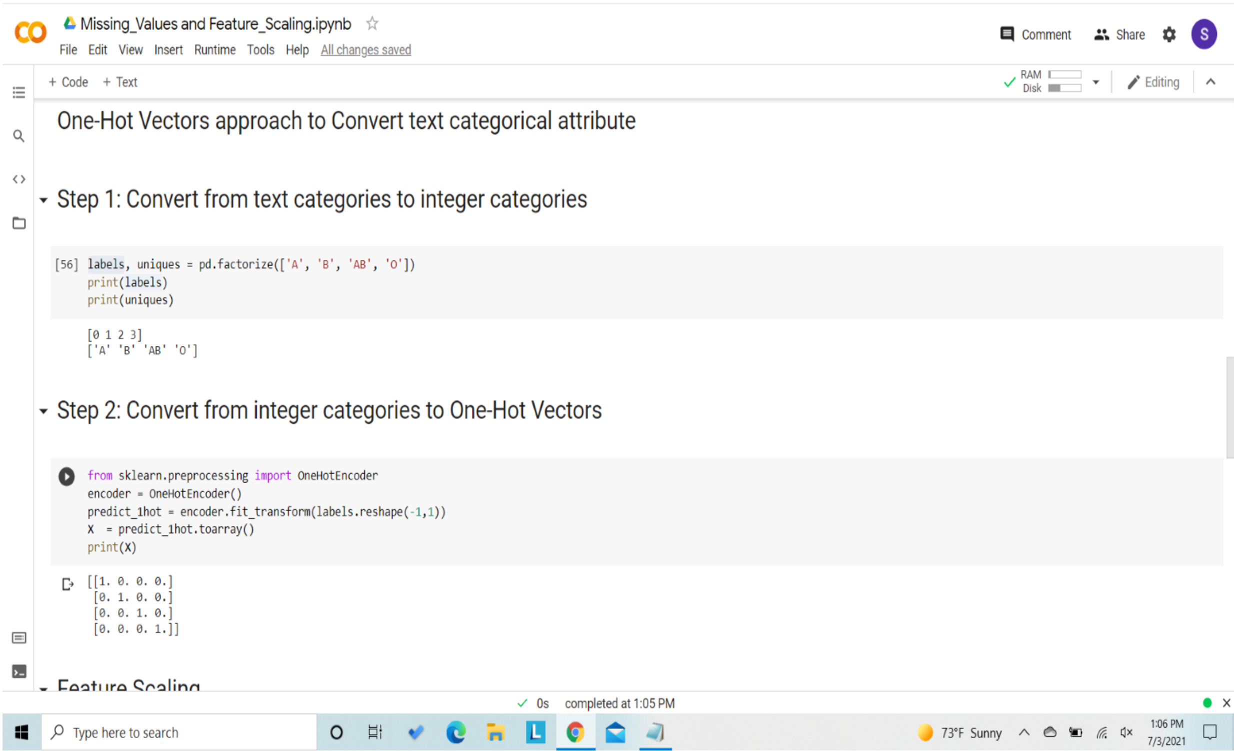
Number of Kids	Working Expereince (years)	Age	Salary	Blood Types
3	15	45	$250,000	A
1	5	30	$200,000	B
2	10	38	$150,000	AB
1	10	36	$180,000	O

Step 1: Convert from text categories “Blood Types” **to integer categories**

Number of Kids	Working Expereince (years)	Age	Salary	Blood Types	Categorical Value
3	15	45	$250,000	A	1
1	5	30	$200,000	B	2
2	10	38	$150,000	AB	3
1	10	36	$180,000	O	4

Step 2: Convert from integer categories to One-Hot Vectors.

Number of Kids	Working Experience (years)	Age	Salary	Blood Types	A	B	AB	O
3	15	45	$250,000	A	1	0	0	0
1	5	30	$200,000	B	0	1	0	0
2	10	38	$150,000	AB	0	0	1	0
1	10	36	$180,000	O	0	0	0	1



4. Please use Standard Scale to scale the data.

Feature Scaling using Standardization.

- Standardization.

Scikit-Learn provides a transformer called StandardScaler for standardization.

Process

- Step 1: It subtracts the mean value (so standardized values always have a zero mean).
- Step 2: It divides by the variance so that the resulting distribution has unit variance.
- Standardization is much less affected by outliers.

1.Input Data -

1	**0**	**0**	**0**
0	**1**	**0**	**0**
0	**0**	**1**	**0**
0	**0**	**0**	**1**

2. Calculate the **mean** for the input data using below Formula -

$$\mu = \frac{1}{N} \sum_{i=1}^{N} (x_i)$$

Here x_i = values of the input data

Note:

- 0.25= (1 + 0 + 0+ 0)/4
- 0.25 = (0 + 1 + 0 + 0)/4
- 0.25 = (0 + 0 +1 + 0)/4
- 0.25 = (0 + 0 + 0 + 1)/4

0.25 0.25 0.25 0.25

3. Calculate the **Standard deviation** using Below Formula -

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2}$$

Here x_i = values of the input data

u = Mean

Note:

- 0.433= sqrt (((1-0.25)^2+(0-0.25)^2+(0-0.25)^2+(0-0.25)^2) /4)
- 0.433= sqrt (((0-0.25)^2+ (1-0.25)^2 +(0-0.25)^2+(0-0.25)^2) /4)
- 0.433= sqrt (((0-0.25)^2+(0-0.25)^2+ (1-0.25)^2+ (0-0.25)^2) /4)
- 0.433= sqrt (((0-0.25)^2+(0-0.25)^2+(0-0.25)^2+ (1-0.25)^2+) /4)

0.433 0.433 0.433 0.433

4. Calculate the **scaled data** using below Formula -

$$z = \frac{x - \mu}{\sigma}$$

X = **values of the input data**

and **u** = **mean** and **σ** = **Standard deviation**

Note:

- 1.732= (1-0.25)/0.433
- -0.577 = (0-0.25)/0.433
- -0.577 = (0-0.25)/0.433
- -0.577 = (0-0.25)/0.433 //Likewise calculate for remaining values

Data after scaling -

1.732	**-0.577**	**-0.577**	**-0.577**
-0.577	**1.732**	**-0.577**	**-0.577**
-0.577	**-0.577**	**1.732**	**-0.577**
-0.577	**-0.577**	**-0.577**	**1.732**

5. Check the **mean** of the scaled data

$$\mu = \frac{1}{N} \sum_{i=1}^{N} (x_i)$$

Note:

- 0.00025 = (1.732 - 0.577 - 0.577 - 0.577)/4
- 0.00025 = (- 0.577 + 1.732 - 0.577 - 0.577)/4
- 0.00025 = (- 0.577 - 0.577 + 1.732 - 0.577)/4
- 0.00025 = (- 0.577 - 0.577 - 0.577 + 1.732)/4

0.00025 0.00025 0.00025 0.00025

6. Check the **standard deviation** of the scaled data

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2}$$

Note:

- 0.99982632866 = sqrt (((1.732 -0.00025)^2+((-0.577-0.00025)^2)*3) /4)
- 0.99982632866 = sqrt (((1.732 -0.00025)^2+((-0.577-0.00025)^2)*3) /4)
- 0.99982632866 = sqrt (((1.732 -0.00025)^2+((-0.577-0.00025)^2)*3) /4)
- 0.99982632866 = sqrt (((1.732 -0.00025)^2+((-0.577-0.00025)^2)*3) /4)

1 1 1 1